Network Task Management Methodologies

Wang Wenbo 24064411G
Zhou Shiyi 24116285G
Zhang Zhihao 24044588G

March 31, 2025

1 Background

Network systems require sophisticated control algorithms to optimize performance across diverse, dynamic conditions. These management approaches have evolved through distinct methodological paradigms, each with characteristic strengths and limitations.

1.1 Rule-based Algorithms

Rule-based algorithms implement handcrafted heuristics to optimize network performance through explicit control rules. Examples include Copa's delay-based congestion control mechanisms and PANDA's bandwidth estimation for video bitrate adaptation. These systems have demonstrated reliable performance in production environments.

However, rule-based approaches fundamentally rely on manual engineering efforts. Network specialists must design, implement, and validate custom rules for each networking scenario. This process becomes increasingly burdensome as network environments grow more complex, limiting adaptability and scalability across diverse operational conditions.

1.2 Learning-based Algorithms

Deep neural networks (DNNs) trained through supervised or reinforcement learning have emerged as powerful alternatives to rule-based systems. These approaches have demonstrated superior performance across multiple networking domains including traffic classification, congestion control, and adaptive bitrate streaming.

Despite their advantages, learning-based solutions face significant challenges:

1.2.1 High Model Engineering Cost

Performance critically depends on specialized DNN architectures tailored to specific networking tasks. This requires substantial expertise in both networking and deep learning. Even when adapting established architectures like Transformers, engineers must manually configure attention mechanisms, tokenization schemes, and other task-specific components. This shifts the burden from rule engineering to model engineering without eliminating the fundamental design overhead.

1.2.2 Low Generalization

Models trained on specific network distributions frequently underperform when deployed in novel environments. For example, ABR models optimized for stable network conditions often fail under dynamic bandwidth fluctuations. This generalization gap undermines operational confidence compared to more predictable rule-based systems, limiting practical deployment despite theoretical performance advantages.

1.3 LLM Method

Recent advancements in large language models (LLMs) suggest potential for developing foundation models for networking tasks. These pre-trained systems with billions of parameters demonstrate emergent capabilities in planning, pattern recognition, and generalization that could benefit networking applications. However, three critical challenges must be addressed:

1.3.1 Large Input Modality Gap

Network monitoring produces structured numerical time-series data fundamentally different from LLMs' native text inputs. This modality mismatch prevents direct application of LLMs to networking tasks without significant input transformation and representation engineering.

1.3.2 Inefficiency of Answer Generation

LLMs generate outputs through sequential token prediction, which introduces:

1. Latency unsuitable for real-time network control
2. Potential hallucination of physically invalid configurations
3. Computational overhead incompatible with resource-constrained network devices

1.3.3 High Adaptation Costs

Bridging the domain gap between natural language and networking requires extensive fine-tuning. For decision-making tasks using reinforcement learning, this process demands prolonged environment interaction that becomes prohibitively expensive given LLMs' large parameter counts. Full-parameter fine-tuning further compounds these resource requirements.

Addressing these challenges represents a critical research direction for leveraging LLMs' capabilities in networking applications while overcoming their fundamental limitations.

2 Innovation

2.1 NetLLM

NetLLM represents the first framework specifically designed to efficiently adapt Large Language Models (LLMs) for networking applications. This innovative approach enables the use of a single LLM (such as Llama2) as a foundation model to tackle various networking tasks without requiring modifications to the base model while achieving enhanced performance. NetLLM consists of three core components:

2.1.1 Multimodal Encoder

The multimodal encoder functions at the input side of the LLM to effectively process diverse input information typical in networking tasks. Its primary goal is to automatically project task inputs from various modalities into the same feature space as language tokens, enabling the LLM to understand and utilize this information for problem-solving. The encoder operates through:

1. Modality-specific feature extractors that process raw inputs from various sources relevant to networking
2. Trainable projection layers that convert these features into token-like embedding vectors compatible with the LLM's input requirements

This approach allows the LLM to seamlessly work with non-textual data types common in networking applications, bridging the modality gap that typically limits LLM applications in technical domains.

2.1.2 Networking Head

To address the inefficiency of traditional token-based answer generation, NetLLM replaces the default language modeling head with specialized networking heads at the output side of the LLM. These networking heads are lightweight trainable projectors that map the LLM's output features directly into task-specific answers.

Rather than generating tokens autoregressively, these heads enable the LLM to produce valid answers in a single inference step by directly outputting from the valid range of possible answers (e.g., selecting a bitrate from candidate options). This approach:

1. Eliminates the risk of hallucination and invalid outputs
2. Significantly reduces generation latency
3. Ensures reliability for time-sensitive networking applications

2.1.3 DD-LRNA

The Data-Driven Low-Rank Networking Adaptation (DD-LRNA) scheme drastically reduces fine-tuning costs while enabling the LLM to acquire domain-specific knowledge for networking. This scheme incorporates:

1. A data-driven adaptation pipeline that works for both prediction and decision-making tasks
2. For decision-making tasks, it employs efficient data-driven reinforcement learning techniques that eliminate time-consuming environment interactions by using an experience pool collected from existing networking algorithms
3. Introduction of additional trainable low-rank matrices (accounting for only 0.31% of total parameters) that efficiently capture networking knowledge

This approach reduces GPU memory requirements by 60.9% and training time by 15.1% compared to full parameter fine-tuning.

2.2 Networking Tasks

2.2.1 Tasks

NetLLM is evaluated on three representative networking tasks that cover different learning paradigms and input modalities:

1. **Adaptive Bitrate Streaming (ABR)**: A reinforcement learning model that dynamically adjusts chunk-level bitrates based on network conditions and playback buffer length during video streaming. The objective is to maximize Quality of Experience (QoE) by balancing factors like chunk bitrate, bitrate fluctuation, and rebuffering time.
2. **Cluster Job Scheduling (CJS)**: Uses reinforcement learning to schedule incoming jobs within a distributed computing cluster. Each job is represented as a directed acyclic graph (DAG) describing execution stage dependencies and resource demands. The scheduler selects which job stage to run next and allocates computing resources to minimize average job completion time.
3. **Viewport Prediction (VP)**: A supervised learning task that predicts a viewer's future viewport positions based on historical viewport data and potentially video content information. This is crucial for immersive video streaming systems (e.g., 360° videos) where only content within the viewer's viewport is streamed in high quality to reduce bandwidth consumption.

2.2.2 Why These Tasks?

These tasks were selected for several strategic reasons:

1. They cover both major learning paradigms commonly used in networking: supervised learning for prediction tasks (VP) and reinforcement learning for decision-making tasks (ABR and CJS).
2. They include both centralized control (CJS, where the scheduler manages the entire cluster) and distributed control (ABR, where clients independently select bitrates) networking scenarios.
3. They involve diverse input modalities that represent the primary data types encountered in networking tasks, including time-series data, graphs, and multimedia content.

VP was specifically chosen as it encompasses multiple input modalities and requires cross-modality fusion, making it more challenging for LLM adaptation than prediction tasks with single input modalities (e.g., bandwidth prediction). Together, these tasks ensure that the evaluation is representative and applicable to a wide range of networking scenarios.

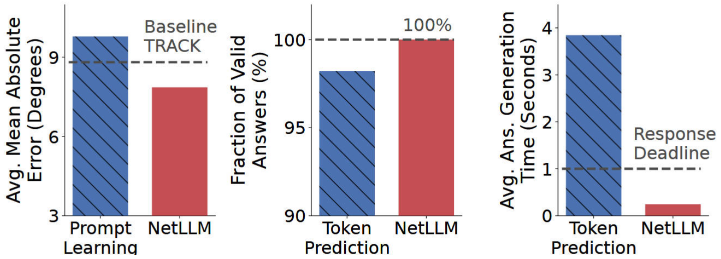

Figure 1. Illustration of the ineffectiveness for some natural alternatives with VP task as the example. Left: Prompt learning that transforms data into textual prompts achieves sub-optimal performance, while NetLLM with a multimodal encoder to encode task input data effectively outperforms baseline. Middle, Right: Token-based prediction with LM head fails to guarantee valid answers and produce stale responses, while NetLLM efficiently addresses these issues with the networking head module.

2.3 Handling LLM Challenges

2.3.1 Large Modality Gap

Networking tasks involve diverse input modalities (time-series data, graphs, images, etc.) that are incompatible with LLMs' text-based input format. This creates a substantial modality gap that prevents direct utilization of LLMs for networking tasks.

While prompt learning approaches that transform data into textual prompts might seem like a solution, they fall short in networking for two reasons: (1) complex modalities like images and graphs cannot be effectively transformed into text, and (2) even when transformation is possible (e.g., time-series data), the performance is suboptimal due to loss of information.

NetLLM addresses this challenge through its multimodal encoder that effectively processes diverse input modalities and projects them into a feature space compatible with the LLM.

2.3.2 Inefficiency of Token-Based Answer Generation

The default token-based generation mechanism of LLMs presents two major limitations for networking applications:

1. The uncertainty in token prediction increases the risk of producing physically invalid answers (hallucination), compromising reliability for critical networking systems.
2. Due to the autoregressive nature of token generation, LLMs often require multiple inference steps to generate a single answer, resulting in high latency that fails to meet the quick responsiveness requirements of networking applications.

NetLLM overcomes these limitations with its networking head that directly maps the LLM's output features to valid, task-specific answers in a single inference step.

2.3.3 High Adaptation Cost

Adapting LLMs for networking tasks, particularly reinforcement learning-based ones like ABR and CJS, involves prohibitive computational costs due to:

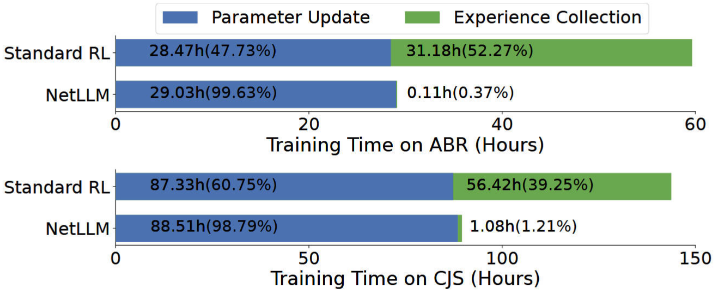

Figure 2. Using standard RL techniques to adapt LLM for RL-based decision-making tasks (ABR and CJS) incurs high training time due to the active environment interaction for experience collection. NetLLM eliminates this time-consuming process by designing an efficient data-driven adaptation pipeline in the DD-LRNA scheme.

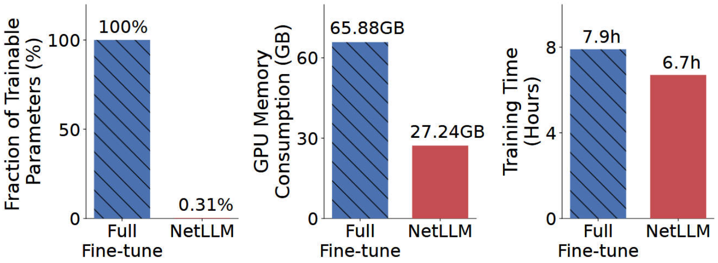

Figure 3. Illustration of the high adaptation costs of full-parameter fine-tune on the VP task. The DD-LRNA scheme of NetLLM efficiently reduces the costs by introducing a set of small trainable low-rank matrices.

1. Time-consuming environment interactions required for experience collection in standard RL approaches
2. High memory and computational requirements for fine-tuning the full parameter set of large models

NetLLM's DD-LRNA scheme addresses these challenges by:

1. Using a data-driven adaptation pipeline that eliminates the need for active environment interaction
2. Introducing efficient low-rank matrices that capture domain knowledge while updating only 0.31% of the model parameters
3. Reducing GPU memory consumption by 60.9% and training time by 15.1%

This approach makes LLM adaptation for networking tasks practical and efficient while maintaining high performance.